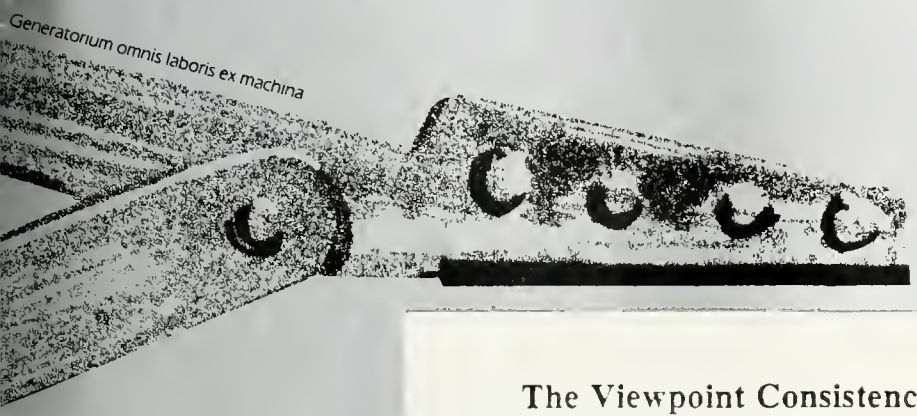
Robotics Research Technical Report

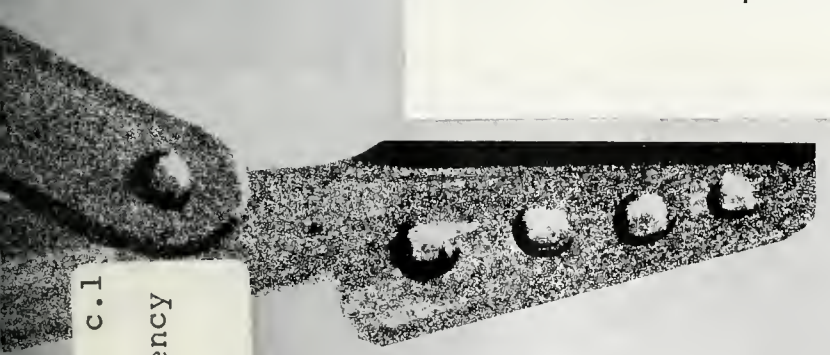


The Viewpoint Consistency Constraint

by

David G. Lowe

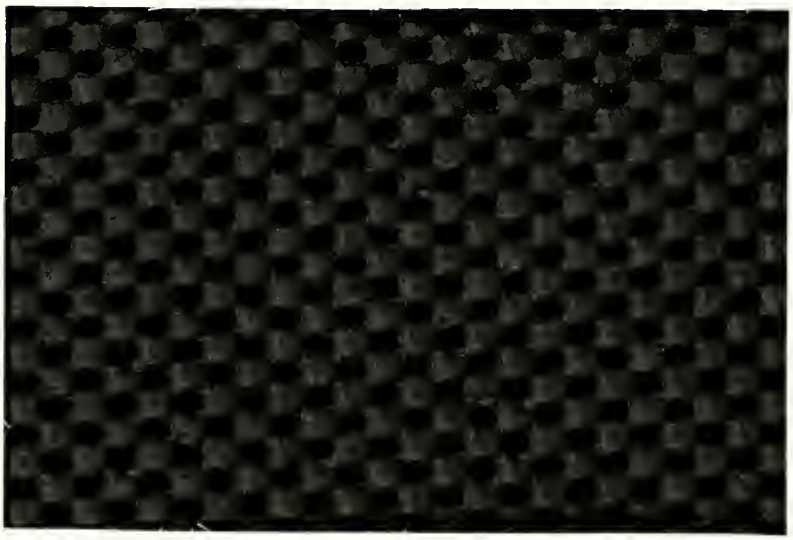
Technical Report No. 244
Robotics Report No. 80
September, 1986



NYU COMPSCI TR-244 c.1
Lowe, David G
The viewpoint consistency constraint.

New York University
Courant Institute of Mathematical Sciences

Computer Science Division
251 Mercer Street New York, N.Y. 10012



The Viewpoint Consistency Constraint

by

David G. Lowe

Technical Report No. 244
Robotics Report No. 80
September, 1986

Courant Institute of Mathematical Sciences
New York University
251 Mercer Street
New York, NY 10012
Arpanet: Lowe@NYU

Work on this paper has been supported by National Science Foundation grant DCR-8502009.

THE VIEWPOINT CONSISTENCY CONSTRAINT

By David G. Lowe

Courant Institute of Mathematical Sciences
New York University
251 Mercer St., New York, NY 10012
Lowe@NYU.arpa

Abstract

The viewpoint consistency constraint requires that the locations of all object features in an image must be consistent with projection from a single viewpoint. The application of this constraint is central to the problem of achieving robust recognition, since it allows the spatial information in an image to be compared with prior knowledge of an object's shape to the full degree of available image resolution. In addition, the constraint greatly reduces the size of the search space during model-based matching by allowing a few initial matches to provide tight constraints for the locations of other model features. Unfortunately, while simple to state, this constraint has seldom been effectively applied in model-based computer vision systems. This paper will review the history of attempts to make use of the viewpoint consistency constraint and will then describe a number of new techniques for applying it to the process of model-based recognition. A method will be presented for probabilistically evaluating new potential matches to extend and refine an initial viewpoint estimate. This evaluation allows the model-based verification process to proceed without the expense of backtracking or search. It will be shown that the effective application of the viewpoint consistency constraint, in conjunction with bottom-up image description based upon principles of perceptual organization, can lead to robust three-dimensional object recognition from single gray-scale images.

Introduction

A fundamental capability of human vision is the ability to robustly recognize objects from partial and locally ambiguous data. As with most problems of interest to artificial intelligence, this high level of performance is achieved through the use of large amounts of domain-specific knowledge, in this case regarding the visual appearance of objects and their components. Methods are known for representing information regarding visual appearance in a computer with a high degree of fidelity, as has been shown by the success of computer graphics in generating realistic images of natural scenes. However, this knowledge itself is of little use without effective methods for applying the constraints implicit in the knowledge during the recognition process.

In this paper we will examine one of the central constraints provided by prior three-dimensional knowledge, which allows us to relate the three-dimensional structure of an object and its components to the two-dimensional spatial structure of its projection in an image. As in other areas of artificial intelligence, the effective application of such a strong constraint leads not only to increased robustness but also to a large reduction in the search space that must be explored during the process of interpretation. The particular constraint that we will be examining can be stated as follows:

> The viewpoint consistency constraint: The locations of all projected model features in an image must be consistent with projection from a single viewpoint.

The ease of stating this constraint is deceptive. The mathematical and practical problems of implementing it have been such that few model-based vision systems have made full use of the constraint. Some systems have ignored it altogether while others have used loose approximations that discard much of the inherent information content. However, the importance of this constraint for achieving robust recognition can hardly be overstated, and we will argue that it plays a central role in most instances of human visual recogntion. Since the appearance of a three-dimensional object can change completely as it is projected from different viewpoints, any attempt to recognize an object without application of the viewpoint consistency constraint will end up ignoring most of the constraining aspects of an object's spatial structure. Low-level vision has proved unsuccessful at generating stable, unambiguous features that in themselves provide reliable discrimination between object classes. However, low-level vision provides not only the identity of features, such as edges, but also accurate information regarding their location in the image. It is this large quantity of accurate spatial information that can be exploited through application of the viewpoint consistency constraint.

A second area of bottom-up image analysis has focussed upon region description, making use of properties such as color and texture. But once again, in themselves these region descriptions are likely to be of little use without a spatial mapping of the object model to the image that specifies which regions are in correspondence to specific surfaces of the object. Thus, spatial correspondence is often prerequisite to other forms of visual matching that are not explicitly spatial themselves.

One argument that is sometimes advanced against the use of precise spatial correspondence is that many objects are non-rigid with internal degrees of freedom and variable dimensions. It is also clear that human vision has a remarkable capability for recognizing

distorted images and drawings. However, advances will be made on these important problems only by explicitly representing the possible degrees of freedom and distortions that are present in a situation. Our knowledge of the visual appearance of objects includes a large amount of information on internal degrees of freedom in their shape and visual properties, as well as potential transformations in the image domain itself. To simply discard all of the available spatial information because some of it is not fully constrained would result in the loss of a large portion of our most useful visual knowledge. It is true, for example, that human vision can identify a person from a highly non-veridical cartoon drawing, yet any amateur artist knows that this is an entirely different proposition from stating that recognition could occur after arbitrary spatial transformations of the image.

The following section of this paper will review the history of the viewpoint consistency constraint in previous computer vision research. Then a particular procedure for its application will be described and examples presented of how the constraint can be applied to limit the amount of search required during visual recognition. This constraint is not only of practical use in current vision systems, but also has important theoretical implications regarding the purpose of the bottom-up components of the visual process. Rather than simply attempting to reconstruct explicit physical properties of the scene, bottom-up vision can also be used to derive stable visual properties of the image that may not have single physical interpretations. Final recognition can be based on a complete mapping of an object model back to the level of the original image, and therefore intermediate bottom-up constructs are of more use for triggering potential matches during the search process than for a final evaluation of the correctness of an interpretation. This leads to an emphasis on such bottom-up processes as perceptual organization which may be more robust in the presence of partial image data than methods that attempt explicit physical reconstruction.

History of the viewpoint consistency constraint

Various forms of the viewpoint consistency constraint have played a role in computer vision research from the earliest days of the field, but the application of the constraint has often been implicit and progress in applying the constraint has been quite uneven. Therefore, it will be instructive to review some of the history of the computer vision field as it relates to the interpretation of spatial information from an image.

Roberts

The seminal work of Roberts [22] in the early 1960's contained many of the important components of spatial interpretation. His vision system began with the detection of edges from a gray-scale image, from which he attempted to form junctions and a graph of connectivities between segments. The interpretation process assumed a domain of rectangular objects, wedges, and pyramids. Based on topological correspondences between parts of the objects and the connectivity graph, sets of matches were hypothesized between points in the image and points on the object. His method for performing spatial verification assumed a particular class of objects and required seven hypothesized point-to-point matches, which could be used to solve for viewpoint and internal size parameters of the

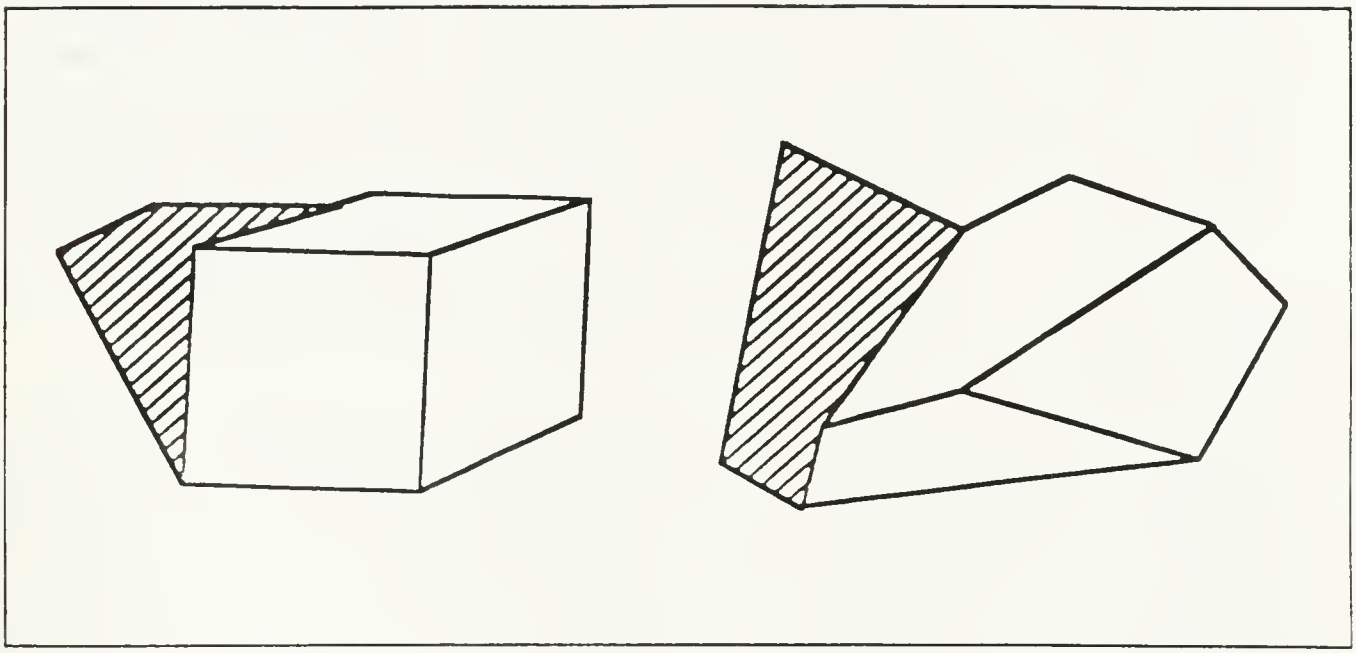


Figure 1: Line labeling is based only on the topological connectivity of edges while ignoring their spatial structure. Therefore, it assigns identical interpretations to the two figures shown above, which have identical patterns of connectivity but different spatial layouts. Only one of these figures is assigned a meaningful three-dimensional interpretation by human vision.

models. The resulting solution was overconstrained and the mean-square error was used to determine acceptance of the match. Although specialized to a restricted domain of models, these methods had many of the robustness properties necessary for the interpretation of real images. It is unfortunate that this work was poorly incorporated in much subsequent computer vision research.

Line labeling

The work on line labeling for polyhedral scenes [3, 12, 13, 25] has often been lumped together with Robert's work because of a similarity in the blocks-world domain in which they were applied. However, the line-labeling methods were based entirely on the topological connectivity of a line drawing and ignored the spatial structure of the image. As shown in Figure 1, while these methods assigned correct interpretations to objects within the specified domain, they would assign identical interpretations to many images that had no plausible physical interpretation. The lack of spatial consistency checks meant that any small error in the input would lead to failure of interpretation. However, there has been some subsequent work by Mackworth [19] and Kanade [14] to add some consistency checks on surface gradients to the basic line-labeling methods.

Pattern recognition

Most of the techniques used in pattern recognition [6] attempt to classify an object from a vector of primitive feature measurements made from the image. Therefore, all spatial information regarding the relative locations of features is likely to be discarded at this

early stage. Since there are seldom stable features that are independent of viewpoint and yet will discriminate between object classes, there is little hope that subsequent statistical operations on this feature vector will result in reliable classification. A similar situation holds for non-quantitative graph matching techniques sometimes used in artificial intelligence, which simply examine adjacency relations or qualitative directional features (e.g., "on-top-of" or "above") while discarding all other spatial information. However, pattern recognition has had some substantial successes in other aspects of the recognition problem, particularly in the important area of learning, so these methods may prove to be very useful if they can somehow be combined with spatial consistency analysis.

ACRONYM

The ACRONYM model-based vision system [1] was perhaps the first attempt to build a complete 3-D framework for incorporating all available spatial constraints during the recognition process. Generic object models could be given to the system in parameterized form with multiple sets of constraints on the individual parameters specifying generic subclasses. Bottom-up processing was based on a search for trapezoid-shaped "ribbons" of certain shapes and sizes predicted from constraints on the model and viewpoint. Actual interpretation was accomplished through a process of narrowing the ranges of unknown parameters, including viewpoint parameters, as each new match was formed. All constraints and measurements were passed in symbolic form to a general constraint module which was expected to return new bounds on the individual unknown parameters reflecting the influence of all constraints. Unfortunately, the solution for general 3-D viewpoint from image measurements involves non-linear constraints whose solution was beyond the capabilities of this general-purpose module. In these cases, it would return bounds that were far from optimal and therefore failed to apply accurate viewpoint consistency constraints. The ACRONYM framework has proved highly influential, but much further work will be required to make effective use of large sets of interrelated constraints.

Goad

A novel method for precomputing viewpoint consistency constraints has been presented by Goad [10] and incorporated in a successful model-based vision system. The method is based on building up tables containing bounds on spatial relationships between selected pairs of image features for small ranges of viewpoints. Recognition is accomplished by a complete search between edges of the model and edges in the image, but as each match is made to a hypothesized image edge its measured location relative to other edges can be checked against the precomputed table to determine additional constraints upon viewpoint. While the initial breadth of this search space is large, after a few matches the viewpoint is constrained to narrow bounds and there is little further search. The precomputation actually solves two separate problems. The most obvious point is that it greatly speeds runtime performance by replacing a complex constraint calculation by table lookup. But in addition, the actual construction of the table can be done entirely in the forward direction by measuring the projected locations of model features from many possible viewpoints. Therefore, at no point does the system need to solve the difficult inverse problem of calculating viewpoint from image measurements. One disadvantage of precom-

putation is that the search path must be largely determined during precomputation, which results in a loss of runtime flexibility. Also, the constraints on viewpoint are only accurate to the size of the parameter ranges for which precomputation was performed, so initial recognition must be followed by a final least-squares parameter estimation [15]. However, for typical industrial problems in which speed of recognition for a few well-specified objects is of most importance, these precomputation methods are likely to remain unsurpassed.

Recognition from depth images

Matching a three-dimensional model to data which is itself three-dimensional can simplify the problem of enforcing spatial consistency. Solutions for solving for the position of a rigid object from matches between surfaces of the object and surfaces detected in depth data have been given by Faugeras [7] and Grimson & Lozano-Pérez [11]. In a different approach, Schwartz & Sharir [23] have developed an efficient procedure for determining optimal matches between subsequences of three-dimensional curves. Depth information can also be used to find stable features for recognition by making use of absolute sizes and angles which would be lost during projection into a two-dimensional image. On the other hand, for human vision the original image input is in the form of a two-dimensional projection, and we have argued elsewhere [17, 18] that most instances of human visual recognition seem to occur prior to the reconstruction of a depth map. Even with carefully engineered sensors, depth data is often time consuming or expensive to obtain. In this paper, we will show that recognition can be reliably performed by determining direct correspondence between a three-dimensional model and a two-dimensional projection and by accurately enforcing the viewpoint consistency constraint. As long as recognition is the goal rather than precise three-dimensional measurement, there will almost certainly be sufficient information in the two-dimensional projection.

Psychophysical studies

As with most aspects of higher-level vision, relatively little is known about the application of viewpoint consistency in human vision. However, some solid psychophysical data is available on the particular topic of mental rotation, which involves the determination of viewpoint parameters which map a prior representation of an object into the coordinates of a particular image. The basic conclusion resulting from numerous experiments is that human vision seems to have a facility for rotating three-dimensional mental models of objects at a fixed rate of rotation. For example, Cooper and Shepard [5] describe an experiment in which subjects first memorized a number of shapes shown at a particular orientation, and were then asked to discriminate these shapes from their mirror-image counterparts after rotation by an arbitrary angle. The time required to perform the discrimination was a linear function of the angular difference between the original orientation during learning and the subsequent orientation during testing. The rotation occurred at a rate of 540 degrees per second, or about eight times faster than in the well-known earlier experiments by Shepard and Metzler [24] in which subjects compared two images presented simultaneously. Nevertheless, mental rotation accounted for up to 30% of the time required to perform the discrimination task, indicating that it can consume substantial computational resources in the brain. Of course, a mature visual system will already be familiar with

the appearance of most common objects from a wide variety of viewpoints, so mental rotation will require negligable amounts of time for typical instances of recognition. The more important requirement may be for accuracy, since accurate viewpoint determination is clearly necessary for many visual tasks, such as determining detailed spatial consistency with a model or judging three-dimensional lengths from a two-dimensional image. Related experiments [2] have shown that other aspects of viewpoint determination, such as size transformation, also happen at a fixed rate.

Enforcing viewpoint consistency

Application of the viewpoint consistency constraint allows us to carry out a quantitative form of spatial reasoning which provides a two-way link between image measurements and the object model. Matches between the model and some image features can be used to constrain the three-dimensional position of the model and its components, which in turn leads to better predictions for the locations of model features in the image, leading to more matches and more constraints.

One component of this problem is the solution for the unknown viewpoint parameters given matches between a three-dimensional model and features in a two-dimensional image. This problem presents a number of mathematical difficulties due to the nonlinear nature of the projection equations. In several previous papers, the author [15, 17, 18] has presented a practical numerical solution to this problem based upon multi-dimensional Newton iteration. This method linearizes the projection equations and uses a novel parameterization to simplify the task of computing partial derivatives of each projected model point with respect to each unknown parameter. Extensions to the basic method have been given for performing least-squares minimization for overdetermined systems, and for minimizing perpendicular distances between lines rather than distances between points. These techniques can also be used to solve for internal model parameters, such as variable sizes and articulations. Each iteration of the Newton convergence requires only a few hundred floating point operations, and convergence to within the accuracy of the data typically requires no more than a couple of iterations. This quadratic rate of convergence is much faster than the linear rate observed during the psychophysical experiments on mental rotation, but it is likely that this difference arises from constraints imposed by the highly parallel architecture of the brain. A number of researchers have explored possible implementations of mental rotation in a parallel network that is consistent with the known psychophysical data [9, 21].

This capability for solving for viewpoint from tentative matches between a model and image features is a prerequisite for application of the viewpoint consistency constraint during the matching process. The numerical implementation of this method is fully described in the references above and will not be repeated here. Instead, this paper will show how this capability can be integrated into the matching process, so that the application of viewpoint consistency can result in a reduced search space during model-based matching. The numerical viewpoint solution techniques allow us to proceed from tentative matches to estimates for the viewpoint or model parameters. In this paper, we will examine the second half of this process—proceeding from viewpoint parameter estimates to

new matches between model features and image features. It is this feedback component of verification that allows the full benefits of a viewpoint consistency analysis to be applied to the matching process.

The essential issue in extending a preliminary match is to allow for robustness with respect to noise and ambiguity in the data. Given the numerical procedures described above, it is straightforward to use a few initial matches to solve for viewpoint and then to project the model onto the image from this viewpoint to predict the locations of further matches. In theory, the task of extending the match would simply require checking for image features at these predicted locations. However, due to noise in the image measurements, errors in modeling, and potential ambiguities arising from closely spaced features in the image, the process of extending a match can easily lead to errors in matching. One solution to this problem would be to make use of a search process, in which a tree of potential matches would be explored through backtracking. However, since the verification process is already on the inner loop of the high-level search for recognition, any search at this level would seriously degrade performance. Instead, we will show how an incremental matching process can be used to greatly decrease the probability of errors during matching, allowing a match to be extended with high confidence and little or no backtracking. This method works by measuring the degree of ambiguity for each match, and selecting only the least ambiguous matches to extend the current set. These new matches are used to update the least-squares estimate for viewpoint, which in turn decreases the ambiguity for the more difficult cases. By the time the most ambiguous cases must be matched, there will usually be a large number of previous matches which provide overconstrained data for a highly accurate viewpoint estimate.

Calculating the probability of false matches

We will assume that an object model has been projected onto the image to provide many predictions for the locations of particular edges in the image. Our task will be to measure the probability of error for matches between particular image edges and edges of the model. Each prediction is assumed to consist of the location and orientation of a straight line segment, and each corresponding image segment may match any subpart of this predicted segment.

We will consider two different sources for these matching errors. One type of error arises from inaccuracies in the prediction which cause an unrelated image edge to appear close to the prediction. We will model this type of error by assuming a random distribution of potentially matching image edges, and then calculating the probability that one could match the prediction to within the measured accuracy. A second type of error arises when there are two closely competing matches in the image, and the wrong one may be selected. This situation frequently arises when related features of an object, such as two closely spaced parallel edges, give rise to edges in the image that have similar locations and orientations. This source of error will be detected by examining all the potential matches for each particular prediction in order to determine ambiguity, from which the probability of selecting the wrong match will be calculated. The two sources of error will be considered for each match and combined to produce a final estimate of probability of error.

Probability of mistaken match to a random segment

Since our initial viewpoint estimate may be based on only a few image measurements, it is quite possible that there will be moderately large errors in the predicted locations of some model features. In addition, projection from three dimensions to two means that features from two unrelated objects may appear arbitrarily close in the image. Since we are unlikely to have much prior information regarding these potential false matches, we will evaluate the probability that a given match could have resulted from some randomly positioned segment. Only if this probability is low can we have confidence in the match. Our calculation of this probability will have much in common with the methods used to detect non-accidental groupings during perceptual organization [18].

We will model the background of false candidates for matching as being uniformly distributed in terms of orientation, position and scale. If more detailed distributional information were available for a particular domain then it could be incorporated, but it is unlikely that such information would be available for natural images taken from arbitrary viewpoints. The expectation of uniform distribution with respect to scale means that changing the size of the image should have no influence on the distribution of segment lengths. Since doubling the size of an image increases the length of each segment by a factor of 2 and the area of the image by a factor of 4, it follows that the density of segments of a particular length per unit area is inversely proportional to the square of their length. Therefore, if d is the density of segments of length greater than l per unit area, then

$$d = \frac{D}{l^2}$$

for some scale-independent constant D. Since the same value of D will appear in all our calculations, the value chosen will have little influence on the ranking of matches. However, for our experiments we have assigned D the value 1, which corresponds to a fairly dense set of segments detected in the image. It is important to take account of the fact that short segments are more common than longer ones, or many false matches would be produced from the large number of short segments that may appear in any part of an image due to texture or noise.

We are now in a position to calculate the expected number of accidental matches, N, that would occur to within a given tolerance of some prediction from among a uniformly distributed set of background segments. When this expected number is much less than 1, it is approximately equal to the probability of the match arising by accident. We will assume that a candidate match has been found whose endpoints lie within the endpoints of the prediction when projected perpendicularly onto the predicted segment. Due to the common occurrence of occlusion and partial failure of edge detection, there is no expectation that a match will cover the full length of the prediction. Let p be the length of the predicted segment, l be the length of a particular matching segment that is being evaluated, and s be the perpendicular separation from the midpoint of the matching segment to the predicted segment (See Figure 2). Then the expected number of matches within the given separation would be given by the density of segments multiplied by a rectangle of length $p - l$ and width $2s$. Therefore,

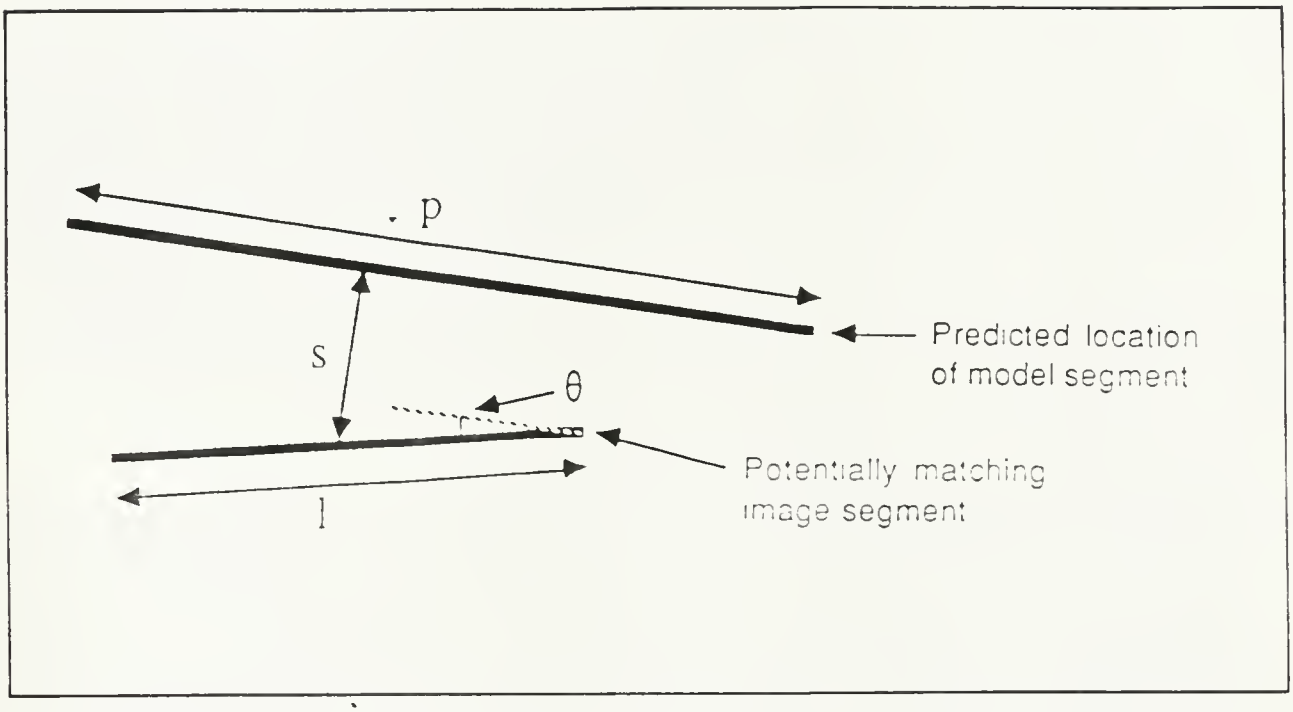


Figure 2: Measurements that are used to calculate the probability of accidental matching between an image segment and a model prediction.

$$N = 2ds(p-l) = \frac{2Ds(p-l)}{l^2}$$

However, we must also take account of the orientation of the matched segment. Let θ be the angular difference in orientation between the predicted and matched segments. Assuming a uniform distribution of orientations, only $2\theta/\pi$ of a set of segments will be within orientation θ of a given prediction. Therefore, after accounting for agreement in orientation,

$$N = \frac{4D\theta s(p-l)}{\pi l^2}$$

This expression provides a measure of the probability of an accidental match occuring to within the specified tolerances in orientation and perpendicular separation.

A separate case occurs when the endpoints of the candidate match do not lie within the endpoints of the prediction. This will occur quite infrequently for correct matches, because it implies some kind of accidental collinearity between the predicted segment and some other continuing segment. The much more common case is for edge detection to find only part of a segment, which is handled by the methods given above. In experiments with matching in real images, we found that this type of extended match occurred less than one-fifth as often as the more normal case. Therefore, the probabilities of accidental matching are multiplied by a penalty factor of 5 for this type of match to decrease their likelihood of being selected.

When these calculations are actually implemented, some care must be taken that realistic values are assigned to all of the measurements. Given the various sources of noise in image measurements, there should be minimum bounds on the measured separations and orientation differences. This prevents an extremely low value for one of the measurements arising by chance or from the effects of discretization and having an undue influence on the final probability estimate. For example, given that the position of a line segment in the image is unlikely to be measured to an accuracy better than 1 pixel, the value of s should not be allowed to fall below a minimum of 1 pixel. Similarly, if the location of each endpoint of a line segment has an error of 1 pixel, then the error in θ will be approximately $2/L$ for a segment of length L (since $sin(a) \approx a$ for small values of a).

Probability of a mistaken match due to ambiguity

A second potential source of mistaken matches arises from situations in which a number of closely spaced parallel lines appear in the image due to the structure of an object, specular reflections, or problems with the edge detection process. Each of these line segments may have a very low probability of having been in close agreement with the prediction by accident. However, since only one match can be correct, there can still be a high probability of making a mistaken match. The solution in this situation is to explicitly measure the ambiguity between competing matches, and to adjust the probability of error upwards when this ambiguity is high.

For each prediction from the model, we evaluate each potential match in the image for the probability that it could arise by accident, using the formulas given above. Let M be the match with the lowest value of this probability for a particular prediction, and $P(M)$ be its probability value. Therefore, if we were to select some match for this prediction, we would choose the match M as the least likely to be mistaken. Now, let $P(N)$ be the next-lowest probability value for the competing potential matches. Clearly, if $P(N) = P(M)$ then we have a 50% chance of selecting the wrong match regardless of the actual value of $P(M)$. More precisely, the probability $P(W)$ of choosing the wrong alternative from among the two best matches is given by

$$P(W) = \frac{P(M)}{P(N) + P(M)}$$

If $P(N)$ is much larger than $P(M)$ then there is little ambiguity and the final probability estimate for making an error will still be small. The value $P(W)$ is calculated for each prediction from the model and is used to select the best matches from among all the predictions to extend the current set of matched features.

Implementation of viewpoint consistency in SCERPO

The methods described above for evaluating and extending preliminary matches play a key role in the implementation of the SCERPO computer vision system. SCERPO is a large computer vision system which combines many components to achieve three-dimensional model-based recognition from single gray-scale images. As with most applications of computer vision to real image data, the lower-level components cannot be expected to function

with high reliability. The methods for extending preliminary matches and enforcing the viewpoint consistency constraint play a vital role by leading to reliable extrapolation and verification of the error-prone matches proposed by the lower-level components. The fact that matches can be extended without backtracking allows the system to perform a significant amount of search within a reasonable budget of computation time. As will be shown, these techniques work well in practice, leading to robust performance under realistic imaging conditions.

Figure 3 shows the various components of the SCERPO system. In the following paragraphs, we will briefly overview the system and place the components for enforcing viewpoint consistency in context. The initial implementation of SCERPO, as described in [17], made use of rather simplistic methods for extending preliminary matches. The improved performance demonstrated in the examples shown in this paper depends largely on the recent incorporation of careful evaluation and incremental extension of matches as described above. Details regarding other aspects of this implementation are being published in a companion paper [18].

The viewpoint consistency constraint is of little use for the initial stages of matching. Since we initially may have no idea of the viewpoint from which we will be viewing an object and may have a library containing large numbers of possible objects, the initial bottom-up stages of vision must detect features that are at least partially invariant with respect to viewpoint and are independent of any specific object. In fact, human vision does have such "perceptual organization" capabilities for detecting bottom-up viewpoint-independent structure in the image. The SCERPO vision system begins by using established methods for edge detection. Figure 4 shows an image of a bin of disposable razors taken at a resolution of 512×512 pixels by an inexpensive vidicon camera. Figure 5 shows the edges detected in this image by finding zero-crossings of a $\nabla^2 G$ convolution [20]. The intensity of each point along a zero-crossing is proportional to the magnitude of the convolution gradient at that point. Straight line segments are detected from these edge points using a scale-invariant segmentation algorithm, producing the set of segments shown in Figure 6. Then a grouping process is executed that detects significant instances of collinearity, endpoint proximity, and parallelism from among these segments. The methods for perceptual organization are beyond the scope of this paper, and the reader is referred to previous papers [16, 17, 18] for a discussion of the derivation and implementation of these grouping properties. The groupings are ranked according to their calculated significance, and some of the most significant groupings that were detected are shown in Figure 7, in which the groupings consist of sets of nearby line segments that share collinearity, parallelism or proximity relationships. The groupings are matched one at a time to components of the object model that are expected to give rise to that type of grouping in the image. In this way, the groupings are used to provide hypothesized matches to trigger the application of the viewpoint consistency constraint.

Matches between an object and the image that are based simply upon viewpoint-invariant properties will necessarily be unreliable. The viewpoint consistency constraint can greatly improve reliability by taking tentative matches between a few image features and object features, solving for a consistent viewpoint, extending the match by predicting the locations of other model features, and iterating. The final decision to accept a match

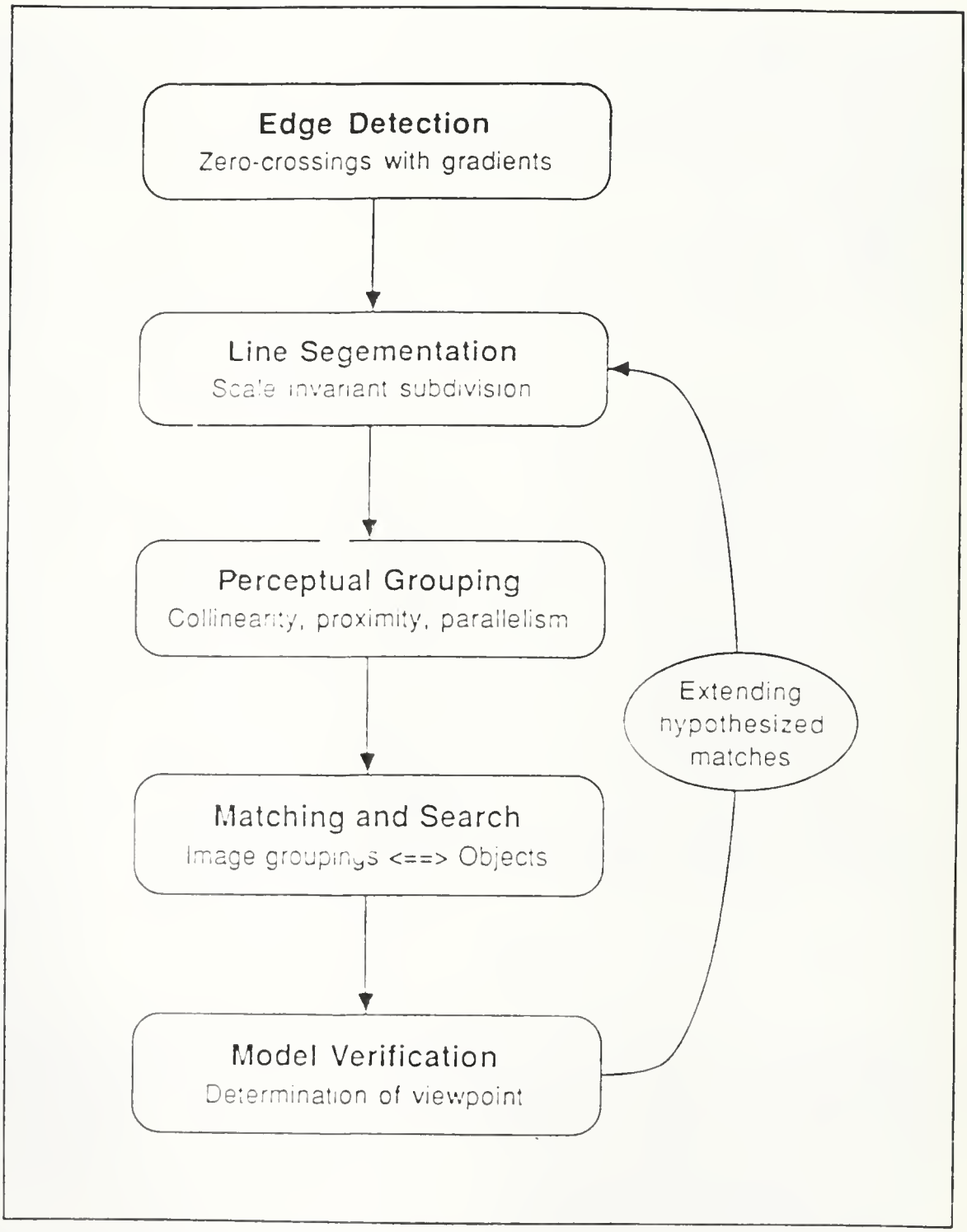
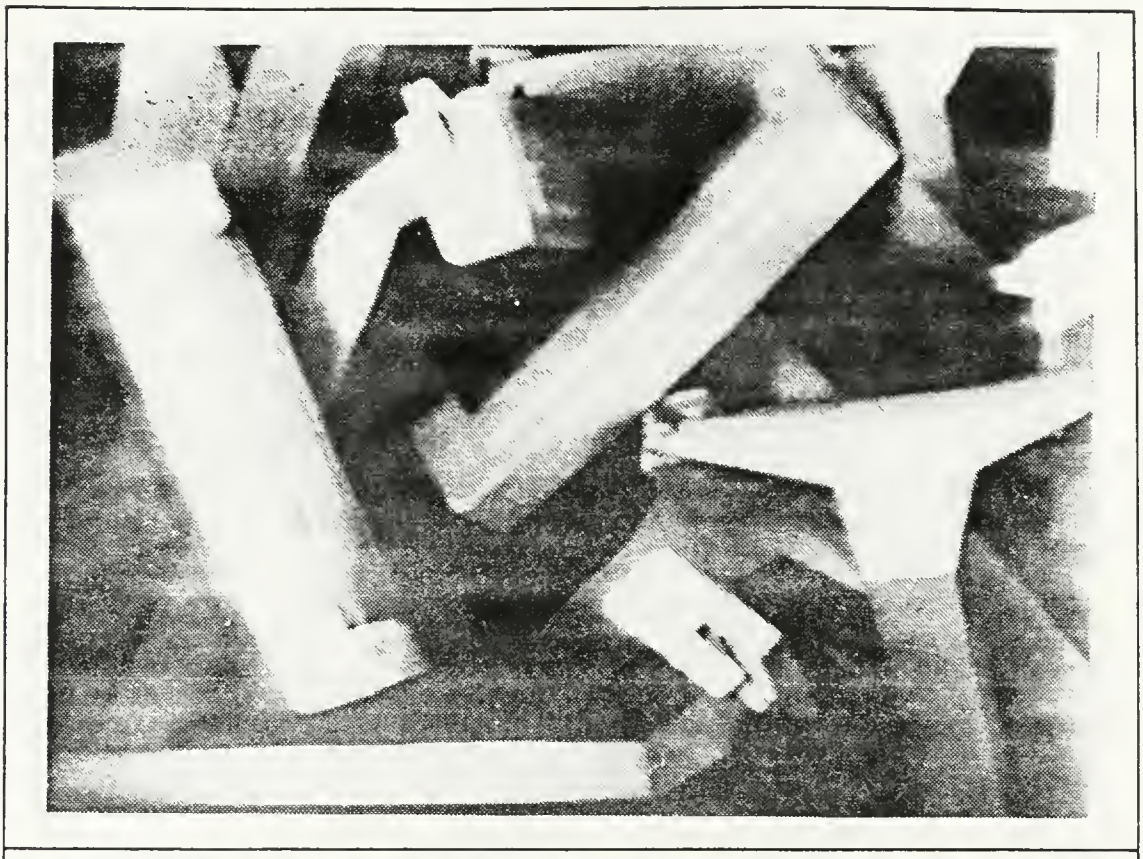


Figure 3: The components of the SCERPO vision system.



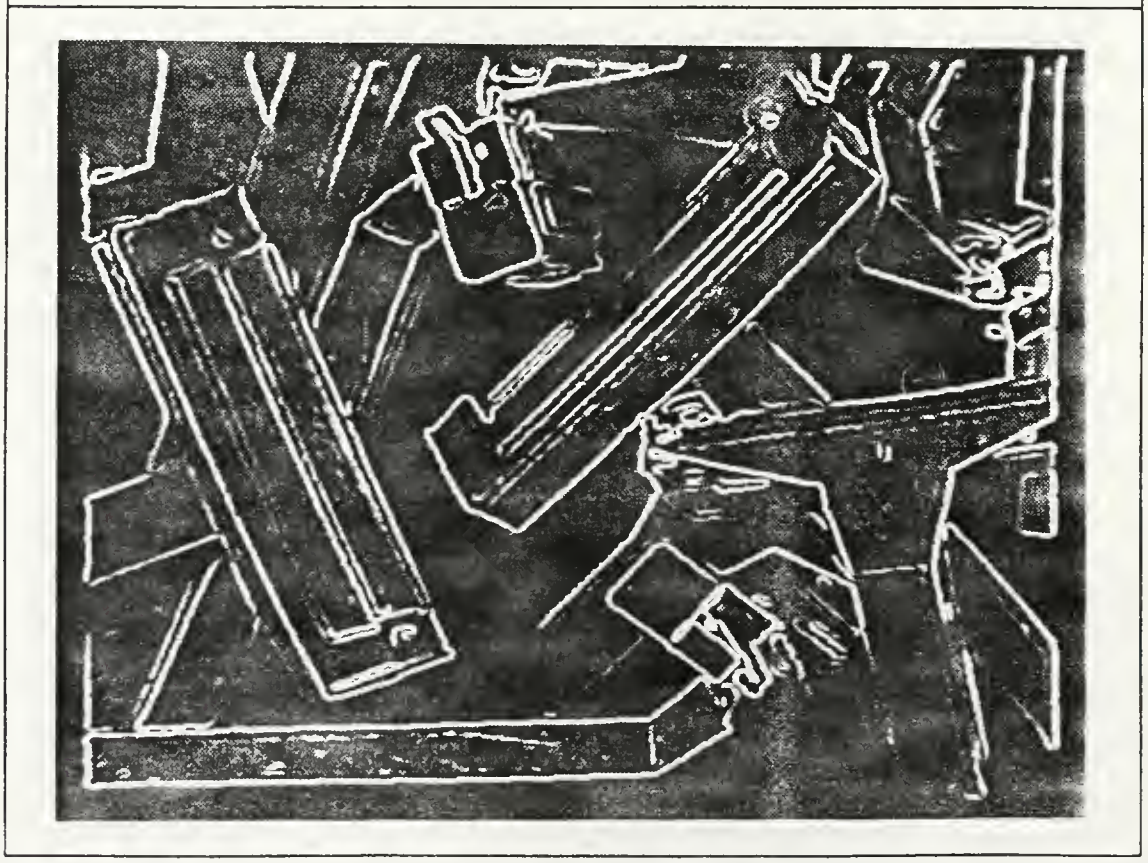


Figure 4: The original image of a bin of disposable razors. **Figure 5:** The zero-crossings of a $\nabla^2 G$ convolution. The intensity is proportional to the magnitude of the convolution gradient.

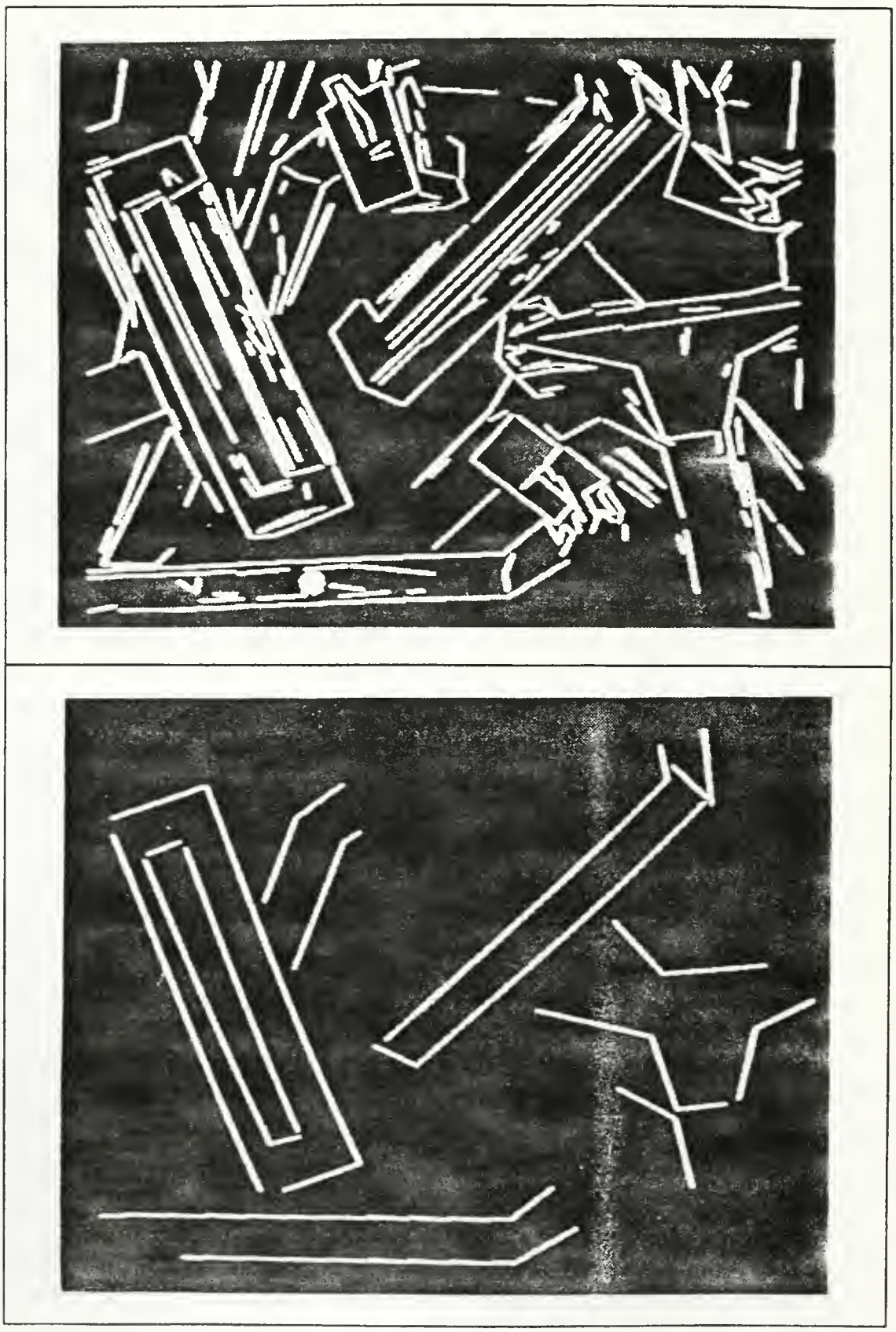


Figure 6: A set of straight line segments derived from the zero-crossings. **Figure 7:** Viewpoint invariant groupings of segments in which each grouping is judged very unlikely to have arisen by accident.

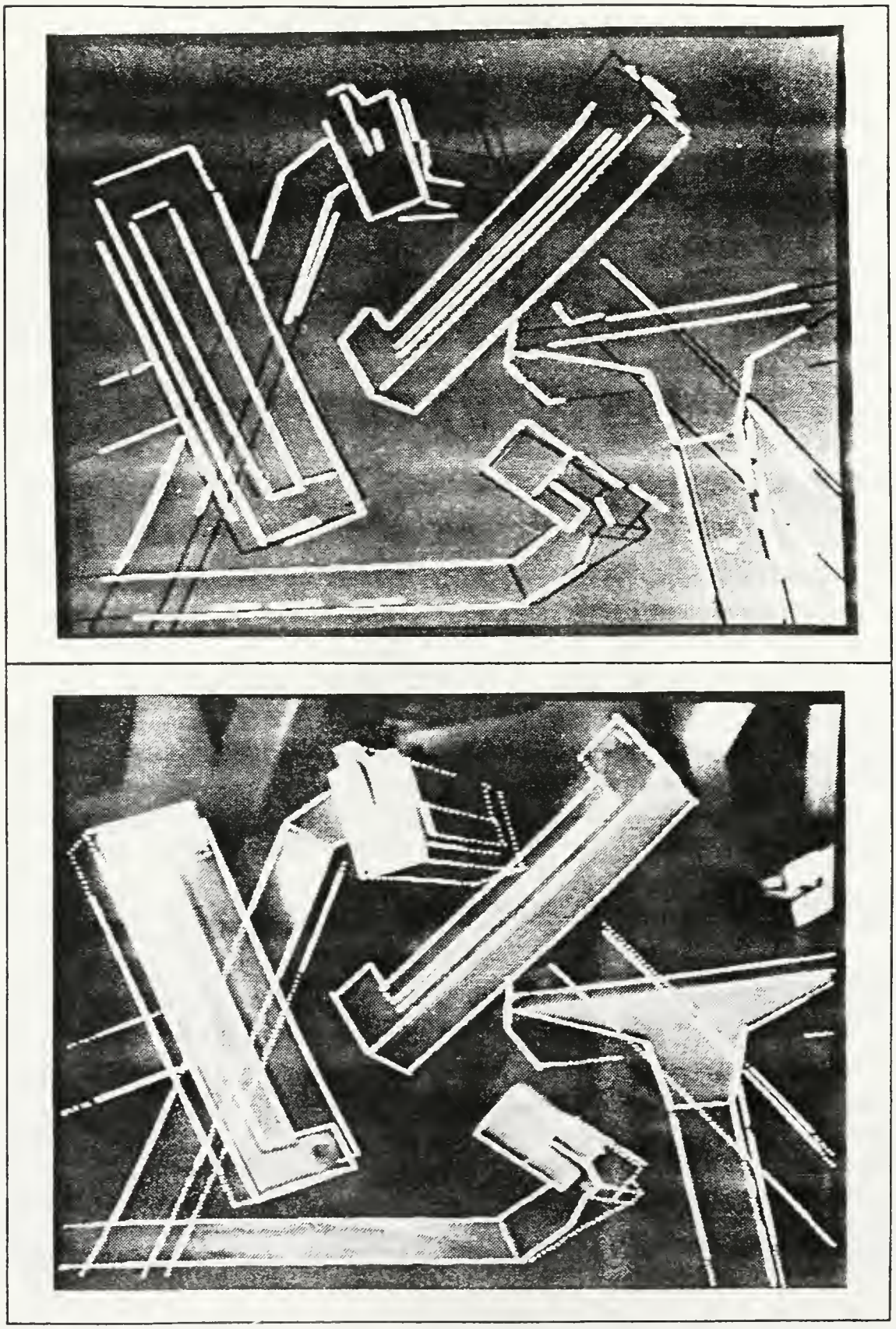


Figure 8: Successful matches between sets of image segments and particular viewpoints of the model. **Figure 9:** The model projected onto the image from the final viewpoints.

is based simply on the number of close matches that were found to a given viewpoint of the model.

As each successful match is found, the identified segments are marked as already matched and are no longer considered for further matching. Therefore, the search space actually decreases as more and more of the segments in the image are removed from consideration. The final results of this process are shown in Figure 8, in which five viewpoints of the model (shown in black) were found to be in close agreement with subsets of the original image segments (shown in white). In each case of successful recognition, more than 15 image segments were matched to the model. Since only about 3 segments are needed to determine viewpoint, all the remaining matches provide confirmation for the presence of the object at that location. Therefore, we can have very reliable identification in spite of partial occlusion and other forms of missing low-level information predicted by the model. Figure 9 shows the model projected onto the image from the final calculated viewpoints. Each edge in this image is drawn solid where there is a matching image segment and is dotted elsewhere. The total computation time expended on this example was about 3 minutes on a VAX 11/785, but it could probably be speeded up by a large factor if speed were a major objective. All of the code beyond the edge detection stage is written in Franz LISP.

Conclusions

Application of the viewpoint consistency constraint greatly simplifies the recognition problem by providing quantitative constraints on the locations of object features in the image. This constraint is strong enough that it can change the basic framework within which recognition is performed. Bottom-up processing need no longer function with high reliability or provide complete representations of physical properties of the scene. Instead, the bottom-up description of an image is aimed at producing viewpoint-invariant groupings of image features that can be judged unlikely to be accidental in origin even in the absence of specific information regarding which objects may be present. These groupings are not used for final identification of objects, but rather serve as "trigger features" to reduce the amount of search that would otherwise be required. Actual identification is based upon the full use of the viewpoint consistency constraint, and maps the object-level data right back to the image level without any need for the intervening grouping constructs.

The matching process presented in this paper is based upon a probabilistic analysis of the likelihood that each potential match is correct. This approach contrasts with the more traditional use of preset error thresholds during matching, which accept any match that is within a range that could be accounted for by noise or modeling inaccuracies. The individual probabilistic analysis of each match can be used to greatly decrease ambiguity and therefore leads to a much smaller search space than would otherwise need to be explored. It is likely that these same methods could be applied to many other components of the recognition problem.

Acknowledgments

This research was supported by NSF grant DCR-8502009. Robert Hummel provided many forms of assistance during the implementation of the SCERPO system.

References

[1] Brooks, Rodney A., "Symbolic reasoning among 3-D models and 2-D images," *Artificial Intelligence,* **17** (1981), 285-348.

[2] Bundesen, Claus and Axel Larsen, "Visual transformation of size," *Journal of Experimental Psychology: Human Perception and Performance,* **1** (1975), 214–220.

[3] Clowes, M.B. "On seeing things," *Artificial Intelligence,* **2** (1971), 79-116.

[4] Conte, S.D. and Carl de Boor, *Elementary Numerical Analysis: An Algorithmic Approach, Third Edition* (New York: McGraw-Hill, 1980).

[5] Cooper, Lynn A., and Roger N. Shepard, "Turning something over in the mind," *Scientific American,* **251**, 6 (December 1984), 106–114.

[6] Duda, R.O. and P. E. Hart, *Pattern Classification and Scene Analysis* (New York: Wiley, 1973).

[7] Faugeras, O.D., "New steps toward a flexible 3-D vision system for robotics," *Proc. of 7th International Conference on Pattern Recognition* (Montreal, 1984), 796-805.

[8] Fischler, Martin A. and Robert C. Bolles, "Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography," *Communications of the ACM,* **24**, 6 (1981), 381-395.

[9] Funt, Brian, "A parallel-process model of mental rotation," *Cognitive Science,* **7** (1983), 67–93.

[10] Goad, Chris, "Special purpose automatic programming for 3D model-based vision," *Proceedings ARPA Image Understanding Workshop* (1983).

[11] Grimson, Eric, and Thomás Lozano-Pérez, "Model-based recognition and localization from sparse range or tactile data," *Int. Journal of Robotics Research,* **3** (1984), 3-35.

[12] Guzman, A., "Decomposition of a visual scene into three-dimensional bodies," *AFIPS Fall Joint Conferences,* **33** (1968), 291-304.

[13] Huffman, D. A., "Impossible objects as nonsense sentences," in R. Meltzer and D. Michie (Eds.), *Machine Intelligence 6* (New York: Elsevier, 1971), 295-323.

[14] Kanade, Takeo, "Recovery of the three-dimensional shape of an object from a single view," *Artificial Intelligence,* **17** (1981), 409-460.

[15] Lowe, David G., "Solving for the parameters of object models from image descriptions," *Proc. ARPA Image Understanding Workshop* (College Park, MD, April 1980), 121–127.

[16] Lowe, David G. and Thomas O. Binford, "Perceptual organization as a basis for visual recognition," *Proceedings of AAAI-83* (Washington, D.C., August 1983), 255–260.

[17] Lowe, David G., *Perceptual Organization and Visual Recognition* (Boston, Mass: Kluwer Academic Publishers, 1985).

[18] Lowe, David G., "Three-dimensional object recognition from single two-dimensional images," *Courant Institute Robotics Report, No. 62,* New York University (February 1986). To appear in *Artificial Intelligence.*

[19] Mackworth, A.K., "Interpreting pictures of polyhedral scenes," *Artificial Intelligence,* **4** (1973), 121-137.

[20] Marr, David, and Ellen Hildreth, "Theory of edge detection," *Proc. Royal Society of London, B,* **207** (1980), 187-217.

[21] Morgan, Michael J., "Mental rotation: A computationally plausible account of transformation through intermediate steps," *Perception,* **12** (1983), 203–211.

[22] Roberts, L.G., "Machine perception of three-dimensional objects," in *Optical and Electro-optical Information Processing,* Tippet *et al.,* Eds. (Cambridge, Mass.: MIT Press, 1966), 159-197.

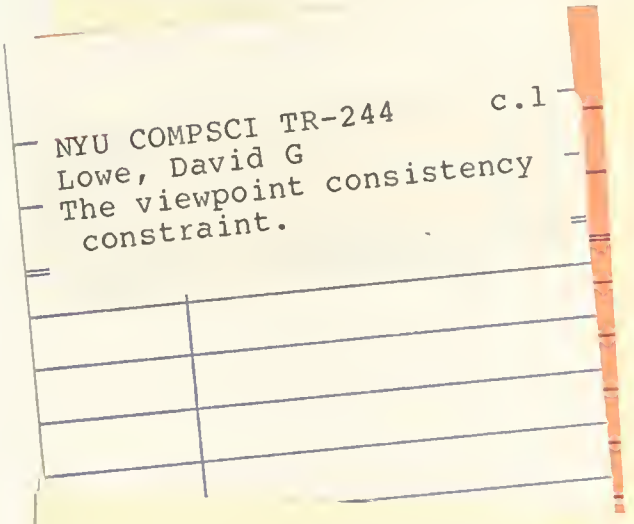
[23] Schwartz, J.T. and M. Sharir, "Identification of partially obscured objects in two dimensions by matching of noisy characteristic curves," *Tech. Report 165, Courant Institute, New York University* (June 1985).

[24] Shepard R. N. and J. Metzler, "Mental rotation of three-dimensional objects," *Science,* **171** (1971), 701-703.

[25] Waltz, D., "Understanding line drawings of scenes with shadows," *The Psychology of Computer Vision,* Ed. P.H. Winston (McGraw-Hill, 1975).

[26] Wolf, Paul R., *Elements of Photogrammetry* (New York: McGraw-Hill, 1983).

NYU COMPSCI TR-244 c.1
Lowe, David G
The viewpoint consistency
constraint.


NYU COMPSCI TR-244 c.1
Lowe, David G
The viewpoint consistency
constraint.

MAR 2 5 1988
DATE DUE

GAYLORD PRINTED IN U.S.A.